Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters

Korkiakangas, Timo

2020

Korkiakangas , T 2020 , ' Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters ' , Studi e Saggi Linguistici , vol. 58 , no. 1 , pp. 67-94 .

http://hdl.handle.net/10138/319140

SSL

Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters

Timo Korkiakangas

Abstract

This paper discusses the theoretical bases as well as the pragmatic implementation of the lemmatization of the Late Latin Charter Treebanks (*LLCT*). *LLCT* is a set of three dependency treebanks (*LLCT1*, *LLCT2*, *LLCT3*) of Early Medieval Latin documentary texts (charters) written in Italy between AD 714 and 1000 (c. 594,000 tokens). The original model for the lemmatization of *LLCT* was the Latin Dependency Treebank (*LDT*), which is mainly Classical standard Latin and based on the entries of Lewis and Short's *Latin Dictionary*. Since *LLCT* reflects later linguistic developments of Latin and contains a plethora of non-standard proper names, particular attention is paid to how non-standard lexemes are lemmatized systematically to make the lemmatization maximally usable. The theoretical underpinnings to manage the lemmatization boil down to two principles: the evolutionary principle and the parsimony principle.

Keywords: treebank, lemmatization, standardization, Medieval Latin charters, onomastics.

1. *Introduction*

Lemmatization: The reduction of the word tokens in a corpus to their lexemes. Thus, the set of word forms or tokens *swim*, *swam*, *swum*, *swims* and *swimming* constitute the lemma for the lexeme SWIM. 'Lemma' is mainly used as an alternative to 'lexeme' or 'headword', the form that heads an entry in a dictionary. (Brown and Miller, 2013: 259)

This paper interprets the above definitions in the way that lexemes are units of lexical meaning while lemma is the form of a lexeme which is conventionally chosen to represent the lexeme. In Latin, noun lemmas are presented in the masculine, neuter, or feminine nominative singular form, depending on the noun's gender; adjectives and pronouns are presented in the masculine nominative singular form. Verbs are given either the present

Received: February 2020
Accepted: April 2020

infinitive form or the first-person singular form of the indicative present. In *LLCT*, the latter form is chosen. With indeclinable parts of speech, the only form is the lemma. Latin lemmatization may look uncontroversial, but things become increasingly complicated as soon as concrete work begins, let alone with non-standard varieties of Latin.

There are currently no generally accepted guidelines for the lemmatization – or the morphological annotation – of Latin. In fact, no publication whatsoever exists that presents a set of principles sufficient for an exhaustive lemmatization or morphological annotation of Latin treebanks, hence the motivation of this special issue. On the one hand, this at first glance surprising defect is possibly motivated by the naïve image, probably fostered by unavoidably restricted normative school teaching, that Latin grammar is straightforward with its exhaustively described, well-defined grammatical categories and transparent lemmas. While this image is not completely distorted within the relatively narrow and well-codified linguistic landscape of Classical Latin, it is plainly untrue for any non-Classical, non-standard variety of Latin. On the other hand, the lack of lemmatization guidelines also seems to arise from the difficulty in systematizing the Latin lexicon satisfactorily, a task that should necessarily be based on extensive lexicographical work. The outcome has been that each project basically follows its own principles of lemmatization and morphological annotation. These principles are typically only described in passing, if at all, in publications on other topics (e.g. Philippart de Foy, 2012; Longrée and Poudat, 2010; McGillivray, 2014). The harmonization of the lemmatization between different Latin resources pursued within the Linking Latin (*LiLa*)[1] project at the Catholic University of Sacred Heart in Milan will no doubt help in establishing a solid ground on which to build a future consensus on Latin lemmatization.

The fluidness of the state of the art is also the reason why the lemmatization of *LLCT* does not form an integral whole. The lemmatization of *LLCT* is a hybrid of various usages adopted pragmatically and, to a certain degree, opportunistically from various sources, mainly from the Latin Dependency Treebank (*LDT*), and supplemented by *ad hoc* practices that looked adequate to manage given non-standard features of charter Latin. A special challenge of *LLCT* is the highly frequent proper names and especially the proper names of Germanic origin with no canonized spelling in Latin. Thus, the aim of this paper is to describe the principles followed in

[1] Cf. *https://lila-erc.eu/#page-top*.

the lemmatization of *LLCT* as exhaustively as possible. The discussion of the lemmatization principles will most often involve the *LLCT* treebanks as a whole (referred to as *LLCT*) while, occasionally, the focus will be on a single treebank (referred to as *LLCT1*, *LLCT2*, and *LLCT3*).

The discussion is organized as follows: Section 2 presents the *LLCT* treebanks while Section 3 briefly characterizes the type of Latin used in charters and defines what is meant by 'standard' in this paper. By giving some numerical data on lemmas in *LLCT*, Section 4 sets the background for Section 5, which discusses the two principles underlying the lemmatization of *LLCT*: the evolutionary principle (Section 5.1) and the parsimony principle (Section 5.2). Section 6 is the conclusion.

2. *The* LLCT *treebanks*

The *LLCT* treebanks consist of three morphologically and syntactically annotated corpora (*LLCT1*, *LLCT2*, *LLCT3*), which also feature a textual annotation layer that indicates abbreviated and restored words. Together the *LLCT* treebanks form a substantial resource for the research of the non-standard non-literary Latin of the Early Middle Ages[2]. Two of the *LLCT* treebanks (*LLCT1* and *LLCT2*) are thus far completed and openly accessible online[3]. The third part, *LLCT3*, is under construction and scheduled to be completed by 2021. *LLCT1* contains 225,834 tokens distributed within 519 charters written in Tuscany between AD 714 and 869, while *LLCT2* contains 257,819 tokens in 521 Tuscan charters from between AD 774 and 897. *LLCT3* will contain ca. 110,400 tokens in 221 charters written in Tuscany as well as in several locations in northern and southern Italy between AD 721 and 1000. The sources of *LLCT1* and *LLCT2* are five copyright-free editions published between 1833 and 1933: Barsocchini (1837), Barsocchini (1841), Bertini (1836), Brunetti (1833), Schiaparelli (1929), and Schiaparelli (1933a). Since most of the charters have also been published recently in the

[2] The other three Latin treebanks are the Latin Dependency Treebanks (*LDT*, *https://perseusdl.github.io/treebank_data/*), the *PROIEL* treebanks (*https://proiel.github.io*), and the Index Thomisticus Treebank (*IT-TB*, *https://itreebank.marginalia.it*).

[3] *LLCT1* is available in Prague Markup Language (*PML*) format at *https://zenodo.org/record/3633607#.XjU4lSNS9EY* and *LLCT2* in *CoNLL* format at *https://zenodo.org/record/3633614#.XjU6zCN7lEY* as well as in the *CoNLL-U* format on the website of the Universal Dependencies consortium at *https://github.com/UniversalDependencies/UD_Latin-LLCT/tree/dev* (see Cecchini *et al.*, 2020).

Chartae Latinae Antiquiores (*ChLA*) series, examples (1) to (6) of the present article will be conveniently referred to by their *ChLA* numbering. For a detailed description of the *LLCT* treebanks, see Korkiakangas (in press)[4].

The syntactic annotation of *LLCT* is based on dependency grammar as operationalized by the *Guidelines for the Syntactic Annotation of Latin Treebanks* (version 1.3; Bamman *et al*., 2007), which, for its part, complies with the annotation style adopted in the Prague Dependency Treebank (Hajič *et al*., 1999). Due to the above-discussed lack of generally accepted guidelines for the morphological annotation or lemmatization of Latin, the lemmatization and morphological annotation of *LLCT1* first practically imitated the choices made in the Latin Dependency Treebanks (*LDT*) available in 2010, the date of the first *LLCT* annotations. The *LDT* lemmas are derived from the Perseus Dynamic Lexicon, which is originally based on Lewis and Short's (1879) *Latin Dictionary* (Bamman and Crane, 2011: 11-13). *LLCT1* was lemmatized and annotated in the Perseus annotation environment, where the Dynamic Lexicon suggested possible lemmas when available. However, it soon became obvious that while the *LDT* style worked for the standard Latin forms of *LLCT*, both a considerable extension of the Perseus Dynamic Lexicon and a set of additional annotation rules were needed to manage the Early Medieval non-standard forms. These rules, described in Korkiakangas and Passarotti (2011), mostly specify principles related to the annotation of morphology, but they also briefly report decisions relative to lemmatization. The same lemmatization practice was originally used with *LLCT2*, which was automatedly annotated and then manually corrected.

The annotation and lemmatization of *LLCT2* were recently thoroughly revised prior to its conversion into the Universal Dependencies style[5]. In its present state, the lemmatization of *LLCT2* can no longer be identified with that of the *LDT* treebanks, based on the Perseus Dynamic Lexicon. At the same time, the possibility of making direct lemma-level comparisons with the *LDT* treebanks is lost. The current lemmatization of *LLCT2* represents a simplified version of the *LDT* style, independent of any predefined lexicon. This style is being utilized for the lemmatization of *LLCT3* as well. In comparison with the newly revised *LLCT2*, the annotation of *LLCT1* looks partly incoherent and should clearly be revised in the future.

[4] For various aspects of the morphological, syntactic, and textual annotation of *LLCT*, see Korkiakangas and Passarotti (2011) and Korkiakangas and Lassila (2013).

[5] The converted version will be distributed in a subsequent release of the Universal Dependencies at the project's website: *https://universaldependencies.org/#language-*.

3. *Early Medieval charter Latin*

Thousands of original Early Medieval charters survive in Italian archives. Charters are legal documents which record private transactions or trials. They were written by quill on parchment by professional or unprofessional lay or ecclesiastical scribes. Charters usually take up one parchment sheet and contain 200 to 1,000 words.

The language of legal documents is always formulaic, and Early Medieval charter formulae draw on a centuries-old legal Latin tradition. However, previous studies suggest that Early Medieval Italian scribes did not copy charters from formulary books, as was done later in the Middle Ages, but had memorized the conventional wordings which they then reproduced with varying success (Amelotti and Costamagna, 1975: 215-216; Schiaparelli, 1933b: 3), hence the considerable linguistic variation. In this way, features of the spoken language, which had evolved far from Classical Latin, occasionally ended up in Early Medieval Italian charters.

Because of this gap between the spoken and written codes, Early Medieval writers had to learn the written code of Latin practically as a second language (Korkiakangas, 2018: 441). Although the gap was wide, the *LLCT* charters suggest that it was still quantitative rather than qualitative. It looks likely that no meta-linguistic split was felt between the spoken language and its written form, both being still considered different sides of one language, Latin. Also, beyond the context of charters, a consciousness of two conceptually different languages seems to have emerged quite slowly in terms of written Latin and spoken Italo-Romance vernacular, a development that eventually led to the first attempts to establish a written form even for the latter (Wright, 2000). The first known reliably datable short texts in the vernacular date from the ninth and tenth centuries, but substantial texts only begin to appear in the following centuries (Frank-Job and Selig, 2016).

Given that Classical Latin standard had to be learnt, the departures from it could be held to be symptoms of the writers' poor school instruction. However, Bartoli Langeli (2006: 25), among others, maintains that, with all its spoken features, charter Latin had established itself as a cherished traditional Italian genre under the Lombard reign («national literature of Lombard Italy»). Be this as it may, charter Latin can be characterized as a 'non-standard' mixture of prefabricated formulae and spoken-language features, where archaic legal terminology is mingled with mistakes and hyper-

corrections provoked by the distance between the sought-after written code and the reality of the spoken language.

At this point, a definition of the term 'standard' (as an opposite of 'non-standard') is needed. In this paper, the 'standard' Latin of the Early Middle Ages refers to a Latin which essentially follows the spelling and morphology of Classical Latin as codified in the prescriptive grammars and used by the Christian authors of the Late Antiquity, who were considered models for literary activity throughout the Early Middle Ages. The spelling and morphology of the Latin of this type show only marginal deviations from those of the Classical Latin of the late Republic and the early Empire while more variation is observed in vocabulary and syntax. This type of standard grammar was still considered the model of written language in Tuscany of the eighth and ninth centuries, judging from other texts of the time as well as from the language of the best *LLCT* scribes. In sum, a rather clear point of reference in terms of a substantial consensus about 'correct' or 'accepted' language use was available in Early Medieval Italy (Korkiakangas, 2017: 577; Bartoli Langeli, 2006: 25 ff.)[6]. However, not all the scribes attained this standard, hence the notable inter-writer variation attested in *LLCT*.

4. *Overall description of the* LLCT1 *and* LLCT2 *lemmatization*

This section provides a background for the following sections by presenting a numerical panorama of the lemmatization of the two parts of *LLCT* already completed, *LLCT1* and *LLCT2*.

Table 1 shows that *LLCT1* contains 4,740 lemmas altogether. The lemma/token ratio is exceptionally low, only 2.1%, which means that each lemma is repeated around fifty times on average. This is because the most common formulae are repeated hundreds of times in the 521 charters of *LLCT1*. 2,139 of the lemmas were available in the Perseus Dynamic Lexicon while the remaining 2,601 lemmas, corresponding to 54.9% of all the lemmas, had to be added manually. 79.8% of the added lemmas were proper names; of all the *LLCT1* lemmas, proper names constitute 49.6%. Moreover, several proper name lemmas only appear once or a few times. These figures reflect well the special nature of charter Latin: many persons involved in the trans-

[6] Cf. Auernheimer's (2003: 49-51) decision to set Alcuin's (essentially Classical) Latin as the point of reference for her study on the Latin of the Carolingian hagiography.

actions are identified, whereas the text proper repeats the same wordings pertinent to its document type (e.g. lease, sales contract, donation) from charter to charter.

LLCT1			*LLCT2*		
tokens	225,834		tokens	257,819	
- lemmas	4,740		- lemmas	3,531	
- of which proper names	2,351	49.6%	- of which proper names	1,860	52.7%
- from *LDT*	2,139	45.1%	- from *LLCT1*	2,428	68.8%
- manually added lemmas	2,601	54.9%	- manually added lemmas	1,103	31.2%
- of which proper names	2,075	79.8%	- of which proper names	805	73.0%
lemma/token ratio	2.1%		lemma/token ratio	1.4%	

Table 1. *Tokens and lemmas in* LLCT1 *and* LLCT2[7].

The overall picture of *LLCT2* is similar to *LLCT1*, although the lemma/token ratio is even lower, 1.4%, with each lemma being repeated over seventy times on average. Such a narrowing is a symptom of the unification of documentary production in the early 9th century, from which the majority of the *LLCT2* charters date. Non-professionals were excluded from notarial practice and establishing chancery traditions entailed a stricter adherence to given formulae (Korkiakangas, 2017: 587; Costambeys, 2013: 246-248), hence the more limited lemma repertoire. *LLCT2* only contains 3,531 lemmas, 2,428 of which (68.8%) were directly transferred from *LLCT1* by way of a simple multi-replace script. For this reason, there is no immediate way to assess to what extent the lemmatization of *LLCT2* coincides with that of *LDT*.

Every corpus of Latin has to decide how to treat certain graphical conventions which change from edition to edition. In the lemmatization of *LLCT*, the character *j* is used before a vowel, whether it was written *j* or *i* in the source edition. Instead, *u* before a vowel is either *u* or *v* depending on the source edition. The *w* of the source editions, attested in words of Germanic origin, is treated inconsistently. In the text of *LLCT1*, it is kept *w* while, in *LLCT2*, it is rendered into the digraph *vu*. The lemmatization utilizes *w* consistently throughout *LLCT*. In *LLCT1*, the traditional Latin convention is followed to capitalize the lemmas that indicate months and calendar

[7] Note that the disambiguation numbers utilized in *LLCT1*, such as *1* in *nomen1* (see Section 5.2), were ignored when calculating the percentages.

terms, such as Kalends, while only proper name lemmas are capitalized in *LLCT2*. *LLCT3* will follow the practices observed in *LLCT2*.

It also needs to be mentioned that *LLCT* uses artificial tokens with no proper lemma to mark gaps in the text (*lacunae*). The artificial tokens are 556 in *LLCT1* and 461 in *LLCT2*. Thanks to the formulaicity of charters, the part of speech of a missing or fragmentary token can often be deduced quite reliably, even without certainty about the exact missing word. In such cases, an artificial placeholder token is created and lemmatized as 'missing^token' in *LLCT2*. For example, in the subscription formula *ego David filio* [Propn] *rogatus* [--] "I, David, son of [Propn], having been asked [--]", a generic [Propn] stands for the proper name expected in that context. It is lemmatized as 'missing^token'. Sometimes, a gap cannot be restored at all, as is the case with the last part of the above example. Then, the artificial placeholder token [--] is used and again lemmatized with 'missing^token'. *LLCT1* is more primitive in its treatment of artificial tokens, which are just marked with '[...]' or '[.....]' and left unlemmatized.

5. *Principles observed in the lemmatization of* LLCT2

The principles presented in the following sections work together in the lemmatization of *LLCT2* and are here separated from each other only for explanatory purposes. The evolutionary principle is presented in Section 5.1, which is further divided into five subsections 5.1.1 to 5.1.5 according to the type of the lemma. Section 5.2 discusses the parsimony principle.

5.1. *Evolutionary principle*

A fundamental principle governing the lemmatization of *LLCT* as well as its morphological annotation is the evolutionary principle which relates the language of *LLCT* to the Classical Latin standard, this latter being understood in the sense explained in Section 3. This principle is also the most distinctive feature of *LLCT* in comparison with treebanks of standard Latin. The evolutionary principle reduces the linguistic variants provoked by language evolution to their standard Latin ancestors. As regards morphological annotation, this reduction sometimes requires an identification of complicated processes which involve both phonological and morphological change in the inflectional ending, whereas with lemmatization, mainly

those evolutionary processes that affect the word stem are concerned. Because word-final inflectional morphemes are used to encode grammatical information in Latin, the evolutionary processes affecting word stems are phonological by nature, with the exception of changes in the number of syllables (see *cuntitigeris* etc. below). Since the challenges related to the lemmatization of proper names partly differ from those related to common names and other parts of speech, the following two sections discuss all other words than proper names, while sections 5.1.3 and 5.1.4 focus on proper names.

5.1.1. *Non-proper-name words with a standard Latin variant*

As regards morphology, the evolutionary reduction of Early Medieval forms to standard Latin forms can be exemplified by the prepositional phrase in (1), where *annus singulus* "every (single) years" is annotated as an accusative plural. This is because the ending *-us* is a typical evolutionary outcome of the standard Latin accusative plural *-os* following the closure of unstressed vowels (Väänänen, 1981: 36). The standard Latin accusative plural is *annos singulos* while the attested *annus singulus* could be misinterpreted, at first sight, as a homonym standard Latin nominative singular *annus singulus*. Obviously, the nominative does not go with a preposition:

(1) *per annus singulus* (*ChLA*[1], 1126)
"every year"

As stated above, with most lemmas it is enough to take phonological evolution into consideration because the morphological change manifests itself principally in inflectional endings. For example, the *LLCT* form *istio* (standard *aestivum*) is lemmatized under *aestivus* "summer-time" (adjective), *anfora* (standard *amphora*) under *amphora*, and *castangneto* (standard *castanetum*) under *castanetum* "chestnut grove". Note that this is done in spite of the fact that forms such as *anfora*, *castangneto*, *presunsere* (standard *praesumpserit*, lemmatized under *praesumo* "to venture"), or *prenda* (standard *prehendat*, lemmatized under *prehendo* "to take"), could very well be lemmatized under their modern Italian successors *anfora*, *castagneto*, *presumo*/*presumere*, and *prendo*/*prendere*, respectively. These fully Italo-Romance forms are likely to have already been in use in the spoken idiom of the time. In other words, the lemmatization of *LLCT* does not seek to describe any particular synchronic stage of Early Medieval Latin. If it did, it should reconstruct contemporary lemmas. That is, however, hardly possible, given the lack of

consensus about Early Medieval spoken Latin. Instead, the lemmatization of *LLCT* seeks to explicate and, subsequently, dissolve the diachronic distance between the attested forms and their standard Latin counterparts in the way that the Latin of *LLCT* is lemmatized as if it were standard Latin[8].

Morphological considerations come into question with lemmas where the stem has undergone alterations in syllabic structure, as is the case with *trentas* (standard *triginta* "thirty") or *poterent* (standard *possent* "they could"). The form *cuntitigeris* seems to be a reduplication inspired by the non-composite stem *tetig-* (standard *contigerit* "he/she may seize"). The evolutionary principle is, however, applied to them in the same way as it is applied to those infrequent cases where a change seems to have taken place in the word formation strategy between standard Latin and Early Medieval Latin: for example, *quattuorcentos* (standard *quadringentos*), lemmatized under *quadringenti* "four hundred" in *LLCT*.

5.1.2. *Non-proper-name words with no standard Latin variant*

The evolutionary principle is relatively easy to observe with Latin-based words discussed in the previous subsection while words that have no standard Latin variant turn out to be problematic. They are often spelled in several different ways, with no binding evidence in favour of one form rather than another. The great majority of the *LLCT* words with no ancestor in standard Latin are nouns, especially proper names (see Section 5.1.4). As for common nouns, words with no obvious standard variant are either loans from other languages, mainly Germanic ones, or Late Latin neologisms. The former include, among others, *sculdahis/sculdais*, a high official under the Lombard reign, *cafagium/gahagias/gahagium* "fenced estate", and *curte/curtis*, which derives from the Greek *khórtos* "courtyard", but seems to have no established Latin spelling. Based on the consultation of the Database of Latin Dictionaries (Brepols)[9] as well as on Nicoletta Francovich Onesti's studies (2000; 2002; 2010) on Germanic loans in Early Medieval Latin and following a careful scrutiny of the word's attestations in *LLCT*, a form that is most likely the common ancestor of the attested forms in terms of its frequency and/or (morpho)phonological features is set to be the lemma. It is either simply picked up among the attested forms or reconstructed if no attested

[8] In the same vein, the morphological annotation of *LLCT* can be used to observe how standard Latin categories are manifested in the Latin of *LLCT*.

[9] Cf. *https://about.brepolis.net/database-of-latin-dictionaries/*.

form seems to represent a (morpho)phonologically plausible ancestor form. In this way, the words above were assigned the lemmas *sculdahis*, *gahagium*, and *curtis*, respectively. As lexicon was not in the core of the projects under which *LLCT1* and *LLCT2* were built, not as much attention was paid to the Germanic words as would have been needed. Therefore, the outcome is often unsatisfactory and sometimes even erroneous in the light of evidence that has turned up during a later consultation of the above-mentioned dictionaries and studies.

Late Latin neologisms are more transparent than Germanic loans. Neologisms can often be assigned, with relative ease, a reconstructed lemma which complies with standard Latin morphology and spelling. This is particularly undisputed when neologisms are derived from standard Latin lexemes by way of usual word formation rules. For example, the adjective *massaricius* "pertinent to a villein holding" and the noun *massarius* "villein, tenant farmer" are regular Early Medieval derivations from the standard *massa* "parcel of land, villein holding" and can be adopted as standard Latin-like lemmas. The same applies to *mustariolum* "wine press", derived from *mustarius* "pertinent to must", or to *patrinius* "stepfather", cf. Italian *patrigno*, originally derived from *pater* "father". In the same vein, standard Latin-like lemmas are coined for less straightforward cases where the derivation involves no affixes and standard Latin models are less frequent: for example, the compound *modilocus* "area which yields one modius", derived from *modius* "corn measure" and *locus* "place, area" (Niermeyer *et al*., 2002, eds.: 911), *reddebeo* "to owe", derived from *reddo* "to pay" and *debeo* "to have to"[10], or the compound pronoun *tumetipse* "you yourself" for *temedipsa* in the phrase *per temedipsa* "by you yourself".

Finally, there are non-derived Early Medieval formations whose origin is not completely transparent: for example, *montone* "sheep" is lemmatized in *LLCT* under *monto*, which seems to be a variant of *multo* "mutton, sheep", cf. Old French *mutun*, modern French *mouton*. Likewise, *sellos* in *sex sellos de olibis* "six measures of olives" is lemmatized under *sellus*, a measure of capacity, possibly originally derived from *situlus* "bucket". If this interpretation is correct, the form postulates a development of the /tul/ group in /l:/ differently from the normal Italo-Romance pattern, where the regular phonological development resulted in /tul/ > /tl/ > /kl/ > /k:j/, like in modern Italian *secchio* (Väänänen, 1981: 65-66); cf. dialectal French *seille*, modern standard French

[10] Cf. Niermeyer *et al.* (2002, eds.: 1169) who use the lemma *redibere*, instead.

seau. Even the meaning of a word may remain unknown, as with *rasula* in the phrase *fini ipsa rasulam de bineam nostras* "up to the *rasula* of our vineyard"[11]. Nevertheless, the form is lemmatized under *rasula*. In this respect, the etymology principle is, in fact, typical of Romance linguistics, which routinely reconstructs 'proto-Romance' ancestors of Romance lexicon.

5.1.3. *Proper names of Latin origin*

As stated above, proper names pose particular challenges to lemmatization in *LLCT*. Since both anthroponyms and toponyms are particularly frequent in charters that record legal transactions between individuals at a certain place and time, a sound treatment of proper names is of the essence in *LLCT*. The challenges are related to two factors, the first of which is specific to *LLCT*: personal names of Germanic origin with no standard Latin ancestors were in fashion in Early Medieval Italy. The lemmatization of the names of Germanic origin involves a number of linguistic problems, which makes them the biggest stumbling block of *LLCT* lemmatization. The other reason is a global one: both anthroponyms and toponyms differ conceptually from common nouns in that their very form has a crucial informational function in identifying the language-external entity to which the name refers.

Proper names are subject to phonological change in the same way as all vocabulary of a given language, but because of their special informational function, they often tend not to be restored to their etymological standard forms in writing even when the writer might have known it, contrary to other vocabulary. As the semantic 'sense' of proper names is subordinate to their 'onymic', i.e. naming, reference (Anderson, 2007: 116 ff.), the etymological roots of names also become forgotten more readily than with normal vocabulary[12]. However, there seems to be a certain gradation in the maintenance of the form of names in *LLCT*, with names of particular importance or familiarity appearing more consistently in a form which was probably commonly felt to be the correct one and which sometimes also involved etymologization, especially if the name had standard Latin models. At least, the names of rulers and of the most important saints testify to such a tendency in *LLCT*, although even they vary quite a lot. On the other hand, the aspiration to restore names to their real or assumed standard Latin forms also

[11] The meaning "abrasion of skin" proposed in Du Cange *et al.* (1883-1887: *s.v. rasula*) does not make sense in this particular context where rather an agricultural term would be expected.

[12] For a detailed discussion on the special features of proper names, see Anderson (2007: § 4).

varies from writer to writer, with a few scribes preferring, for example, the hypercorrect *Latiarus* to *Lazarus* and *Austripertus* to *Ostripertus*.

In general, those proper names that have ancestors in standard Latin are lemmatized following the etymology principle as explained in Section 5.1.1. This is uncontroversial in transparent cases, such as *Pretestatus* (lemmatized under *Praetextatus*), *Deusdede* (lemmatized under *Deusdedit*), originally Greek *Aeleutieri* (lemmatized under *Eleutherius*), or toponym *Ilice* (lemmatized under *Ilex*). However, it is sometimes difficult to decide whether certain names, such as *Liliodarus*/*Lilioderus* or *Theopingtus*/*Thepingtus*, originally have ancestors in standard Latin or whether they are rather combinations of Latin and Germanic elements, like, for example, *Clarisinda* clearly seems to be. *Liliodarus* and *Lilioderus* are lemmatized under *Liliodorus* and may be originally composed of *lilium* "lily" and *dōron* "gift", a typical element of Greek anthroponyms. *Lilio-* is also attested in other *LLCT* names, such as *Liliaufunsus* (lemmatized under *Liliofonsus*), *Liliopinctus*, and *Liliolus*. *Theopingtus* and *Thepingtus* are lemmatized under *Theopinctus*. On the one hand, the name could be a variant of the late Greek *Theópemptos* or *Theópentos* while, on the other, *pinctus* may mean "decorated, adorned", from *pingo* "to paint", a meaning that would make sense in *Liliopinctus*; cf. Italian compounds, such as *variopinto* "multicolour". The first element of *Theopingtus*/*Thepingtus* can also be inspired by Germanic names, such as *Teutfrid* and *Teopaldo*, which begin with the popular Germanic element *t*(*h*)*eu-*/*t*(*h*)*eo-* (< **þeuð-* "tribe, people") (Francovich Onesti, 2000: 216; Francovich Onesti, 2002: 1142).

With some undoubtedly Latin-based names, it is not obvious what the original form is, as phonological development has obscured it and several close variants may occur side by side. This situation is typical of toponyms. For example, it can be duly asked whether the forms *Rocta*, *Ropta*, *Rotta*, and *Rota* are different spelling variants of the same toponym. The first three quite likely derive from the standard Latin participle *rupta* "broken, i.e. rocky", while the last one could equally well come from *rota* "wheel". Based on topographical considerations, they are all lemmatized under *Rupta*.

Any uncertainty about the standard Latin ancestor form of names that only occur in one form in *LLCT* leads to the sole attested form being taken up as the lemma: for example, the toponym *Coltserra* or the anthroponym *Inquircius*. As the *LLCT* treebanks were lemmatized over a long period of time, new instances kept turning up over the process that called for a reappraisal of the previously assigned lemma. The lemmatization has sometimes

failed to be changed accordingly, a fact that contributes to the present incoherent state of the lemmatization of proper names in *LLCT*. Moreover, a deliberate differentiation is sometimes applied in cases where there is insufficient proof to identify two or more slightly differently spelled anthroponyms or toponyms with each other. For example, it is not sure that *Sarturiano* and *Satoiano* (lemmatized under *Sartorianum* and *Satoianum*, respectively) refer to the same place even though that seems possible on phonological grounds. All this having been said, there is no doubt that a scrupulous onomastic revision would radically improve the lemmatization of *LLCT*. As mentioned above, the reason behind the present deficiencies in the lemmatization of proper names is that onomastics did not rank among the interests that guided the building of the *LLCT* treebanks, where the focus has always been on morphology and syntax rather than vocabulary.

Sometimes, it is not clear whether a second-declension toponym that ends in *-o* should be interpreted as neuter or masculine. This is because the neuter as an independent gender category had practically disappeared by the Early Middle Ages and because the *-o* ending can be argued to represent the Romance-type default form of the singular *-o* declension, derived from the accusative in *-u*(*m*) (for both masculine and neuter; Smith, 2011: 278, with references; Korkiakangas, 2016a: 291-295; Korkiakangas, 2016b: 72-73). It was decided that with toponyms ending in *-o*, the *LLCT* lemma ends in *-um* if it is not clearly based on a certain unquestionably reconstruable form of other gender, as is the case with *Saltucclo*, which must be derived from the masculine noun **saltuculus* (diminutive of *saltus* "forest") and is lemmatized as such (*Saltuculus*). For example, the toponym *Sexto* (modern *Sesto*) in *de loco Sexto* "of the place Sexto" and in *ad Sesto* is lemmatized under the neuter noun *Sextum*, although it could also be lemmatized under the masculine adjective *Sextus*, especially when it occurs with *loco* "place". However, in most cases, the elliptical *loco* construction cannot be used as a proof because it allows lack of agreement: for example, the feminine noun in *in loco Valeriana* and the genitive in *in loco Capelle*. Regrettably, an opposite decision was made concerning those third-declension toponyms whose gender cannot be deduced from the form attested in *LLCT*, such as *Lunise* in *ad Lunise* or *Montise* in *ubi dicitur Montise* "which is called Montise". They were interpreted as masculine accusative forms and assigned the masculine lemmas *Lunensis* and *Montensis*, respectively, despite the fact that the forms in question could be neuter (or feminine) accusatives as well. Third-declension toponyms of this kind are infrequent, though.

5.1.4. *Proper names of Germanic origin*

As was suggested above, the lemmatization of proper names of Germanic origin is even less accurate and less coherent than that of Latin-based (or originally Greek-based) names. Therefore, it is not recommended to use the lemmatization of *LLCT* for onomastic investigations.

The evolutionary principle cannot usually be sensibly applied to the Germanic names that occur in *LLCT* because they almost never have obvious standard variants. The cases closest to a standardization of any kind include rulers' names, such as *Carolus*/*Karolus* or *Berengario*, lemmatized under *Carolus* and *Berengarius*, respectively. As a rule, each name has to be evaluated separately based on research on historical Germanic languages. In this respect, the studies of Francovich Onesti (2000; 2002; 2010) have again been of great help, but, as stated above, they were not consulted in a systematic way under the construction phase of the *LLCT* treebanks. Moreover, knowledge on original Germanic morphological elements only helps in recognizing them behind Early Medieval Latin names and, thus, in unifying the spelling of that element in the lemmatization. Occasionally, it also helps in matching two very differently spelled names under one lemma.

However, Germanic morphology results in highly varying outcomes in the Latin of charters. For example, according to Francovich Onesti (2000: 173), the element **agjō* "blade" can be recognized in charters behind the elements *Agi-*, *Aghy-*, *Age-*, *Atge-*, *Ag-*, *Agg-*, *Agel-*, *Agil-*, *Achi-*, *Ahci-*, *Aci-*, *Ace-*, *Ac-*, *Acu-*, and *Ai-*. Yet, some Germanic-based onomastic elements seem to represent established Tuscan types: for example, the spellings *Achi-* and *Agi-* are particularly frequent in *LLCT*. Thus, even though it might be possible in some cases, it is of no use to seek to reduce the immense spelling variation conditioned by Early Medieval Latin phonology to any artificial Germanic lemma by creating lemmas beginning with *Agjo-* for this specific morpheme (e.g. *Agipert* lemmatized under fictitious *Agjoberhtaz*). Instead, it is possible to recognize whether a certain linguistically plausible form is clearly a preferred one in terms of frequency and then to use it as the lemma. Alternatively, the considerations on frequency and Germanic morphology may help reconstruct a lemma as the common denominator to all the attested forms. In spite of this, decisions have been difficult, and, for example, the forms *Agiulo*/*Aggioli* (genitive), *Agguli* (genitive), *Aculo*, and *Aiuli* (genitive) have ended up with four lemmas in *LLCT*, *Agiolus*, *Aggulus*, *Aculus*, and *Aiolus*, respectively, although there seems to be no reason not to consider them representatives of the same lemma, whatever that might be (perhaps *Agiolus*). Although the ap-

plication of the etymology principle is reduced with Germanic names, special care was taken to ensure that names that refer to a certain person are always lemmatized under one lemma. For example, *Hluttarius*, *Hlotharii* (genitive), and *Lotharii* (genitive), all referring to the king Lothar, are lemmatized under *Hlotharius*. The same applies to notaries or other identifiable persons that occur a number of times in one or in several charters. Further, lemmatization is sometimes inconsistent between *LLCT1* and *LLCT2*: for example, *Ildicari* (genitive) and *Ildechieri* (genitive) have mistakenly ended up with two lemmas, *Ildicarus* in *LLCT1* and *Ildecherus* in *LLCT2*.

The Germanic-based masculine names of *LLCT* appear either with Latin inflectional endings, with the Germanic ending *-i* (Francovich Onesti, 2000: 233), or without inflectional endings at all: for example, *Gunfridus*, *Gunfridi*, and *Gunfrid* are all attested. The choice between the three seems to be idiosyncratic, but the Latin endings are by far the most frequent. All the feminine names end in *-a* in the nominative singular (e.g. *Aliperga*) while the names that have entered the Latin third declension (e.g. *Frido*) are usually inflected according to the nasal paradigm (e.g. *Friduni*, dative; Francovich Onesti, 2000: 240) and are, consequently, easy to lemmatize (*Frido*). The *LLCT* lemmatization adds the Latin inflectional ending *-us* to those names that have entered the Latin second declension at least once in *LLCT*; for example, the above *Gunfridus*, *Gunfridi*, and *Gunfrid* are lemmatized under *Gumfridus*. Quite rare Germanic names, such as *Aloin*/*Aloni* or *Eoin*, never appear inflected in *LLCT*, hence their lemmatization without inflectional endings (*Aloin* and *Eoin*, respectively). This practice is identical with the one observed with Biblical names that are traditionally used uninflected and are lemmatized accordingly (e.g. *Daniel*, *Abraham*). Yet other names fluctuate between the second and third Latin declensions, which has sometimes led to inconsistent lemmatization decisions: the nominative and genitive form *Waltari* gets an accusative form *Uualtarene* and is lemmatized under *Waltarus*, although the genitive *Waltari* does not necessarily entail belonging to the second declension.

5.1.5. *Evolutionary principle with mistaken expressions*

This section discusses the import of the evolutionary principle on the lemmatization of mistaken words in *LLCT*. Such a scenario is irrelevant with literary corpora, where erroneous forms are not present, but is pertinent with charters, which are unemended original documents and feature significant linguistic irregularities.

Let us first consider how the evolutionary principle is applied to erroneous morphosyntax. In order to cope with the non-standard morphology of *LLCT*, Korkiakangas and Passarotti (2011: 106 ff.) coined an annotation principle based on 'functional' and 'formal' analyses of morphosyntax. The principle operates on the syntax/semantics interface, linking attested morphological forms to their standard Latin ancestors with the help of the evolutionary principle. Importantly, it also deals with erroneous forms that are impossible from the viewpoint of language evolution, i.e. motivated extra-linguistically. In such cases, an attested morphological form does not match with its expected standard Latin function on the syntax/semantics interface. For example, in (2), the coordinated ablative/dative form subject *heredibus nostris* "our heirs" depends on the predicate *habitare debeamus* "have to dwell".

(2) *Tam nos quam et heredibus nostris in ipsa casa habitare debeamus.* (*ChLA*[1], 1061)
"Both we and our heirs have to dwell in that house."

In standard Latin, the subject of the finite verb is always marked with the nominative case. The form *heredibus nostris* cannot be a morphophonological evolutionary outcome of the standard Latin nominative form *heredes*, and, therefore, it cannot be marked functionally as a nominative. *Heredibus nostris* must be a linguistic error due to a contamination between two or more formulae, a phenomenon frequent in charters, or to an infelicitous interpretation of the abbreviation *hhd* (for *heredes*) (Korkiakangas and Passarotti, 2011: 107). In *LLCT*, functionally impossible mistaken forms of this kind are simply assigned a formal morphological analysis that corresponds to the evolutionary ancestor of that form in standard Latin. Thus, *heredibus nostris* receives an ablative/dative plural morph tag although the subjects of finite verbs cannot be marked with such a case in any variety of Latin.

While the practice described above is fundamental to the annotation of non-standard morphology, it also plays a marginal role in lemmatization, where the question is basically about semantics. Words that are incongruent, i.e. mistaken, in their present context are found sporadically in *LLCT*. In literary texts, one is not accustomed to find mistaken words because literary texts are transmitted through centuries of copying and emendation and finally subjected to editing based on textual criticism. Instead, the scribes who wrote charters were not always equal to their tasks in this respect. Some

misunderstood expressions, usually in age-old documentary formulae, are characteristic of a single scribe, while others are used by more scribes, suggesting thus a local convention. For example, a few scribes mistakenly use in (3) the form *genium*, which looks like an accusative singular form of the word *genius* "tutelar deity, genius", in lieu of *ingenium* "natural disposition, machination, scheme"[13], a word that normally appears in the formula of (3) and makes sense in that context. The translation of (3) conveys the intended meaning (*ingenium*).

(3) *Si forsitans quicumquem de heredis meis* [...] *substraheret quesieret per colive genium*. (*ChLA*[1], 1058)
"If anyone of my heirs [...] perchance tries to dispossess [something] by whatever scheme."

Genium is not an evolutionary outcome of any morphophonological process of *ingenium*, but a blatant misinterpretation resulting from the writer having confused *ingenium* with *genius*, the latter most likely absent in the spoken vernacular of the time. In (3), *genium* is lemmatized under *genius*, which is the only possible standard Latin source for the attested form. This kind of lemmatization follows the practice of formal analysis observed with non-standard morphology and illustrates the uncompromising mode of operation of the evolutionary principle: it always reduces an attested form to its morphophonologically possible language-evolutionary ancestor, whether it makes sense or not in terms of the integrity of the construction or its meaning.

As stated, clearly mistaken words are relatively infrequent in *LLCT*. Additionally, with most mistakes, the formal analysis is obvious and the application of the etymology principle banal: this is the case if the attested word is completely different from the expected/intended one, such as *tradedimus* "(we) handed over/commissioned" in (4), where *rogavimus* "(we) asked" would have been expected on the basis of numerous occurrences. The translation again conveys the intended meaning (*rogavimus*).

(4) *Quam biro cartolas binditionis nostres ad nus factas Warnegausu notarium iscriberes tradedimus*. (*ChLA*[1], 732)
"We asked the notary Warnegausu to write these sales contracts which we made."

[13] Du Cange *et al.* (1883-1887: *s.v. ingenium*).

Tradedimus is not etymologically derived from *rogavimus*, which normally appears in this formula, and is lemmatized formally under *trado* "to hand over/to commission". The writer has probably confused the construction with *trado* with a gerund, which is, however, only attested once in charters (sentence in (5)). Here, the gerund is *scriuendo* "to be written" while the sentence in (4) shows an infinitive (*iscriberes*, i.e. *scribere*).

(5) *Ego Uualtprand in Dei nomine episcopus in hanc cartula donationis* [...] *manus meas suscribsi et confirma et scriuendo tradedi.* (*ChLA*[1], 911)
"I, Waltprand, bishop in God's name, subscribed [...] in this donation and confirmed [it] and commissioned [it] to be written."

In conclusion, it must be stated that the lemmatization of mistaken expressions in *LLCT* has not been as systematic as would be desired. In the sentence in (6), the writer has written *insunt* "(they) are in" instead of *hi sunt* "these are". The former is a nonsensical misinterpretation of the latter, which is the normal way to introduce a list of names in the formula in question and a variant of the frequent *id est* "i.e.". However, when the sentence was lemmatized for *LLCT2*, *insunt* was 'normalized' by splitting it into two tokens, and *in* lemmatized as *hi* "these" under *hic* "this" and *sunt* "(they) are" under *sum* "to be".

(6) *Direxistis missos tuos, in sunt Petrus notario de Uuamo et Sicholfo.* (*ChLA*[2], 85, 37)
"You sent your envoys, they are Petrus, the notary from Guamo, and Sicholfo."

That the writer has written *insunt* intentionally is proved by the following sentence, which lists another set of envoys and also features *insunt*. Thus, to be consistent with the practice described in this section and to respect the choices made by the charter scribes, *insunt* should be restored as one token and lemmatized under *insum* "to be in" as soon as *LLCT2* is again revised some day in the future.

5.2. *The Parsimony principle and homonymous lemmas*

The other general principle that is applied to the lemmatization of *LLCT* together with the evolutionary principle can be called 'the parsimony

principle'. This means that the lemmatization style of *LLCT* does not seek to multiply lemmas unnecessarily. As stated above, not only spelling, but also inflectional morphology fluctuated in Early Medieval Latin. One solution to cope with forms that have changed their inflectional properties is to provide these non-standard forms with new lemmas. This is what some dictionaries do when they provide separate entries to pre-Classical gender variants, such as *corium* (neuter) as opposed to *corius* (masculine) "skin" (e.g. Forcellini *et al.*, 1858-1875; Gaffiot, 1934). Such a solution does not, however, do justice to later written Latin, where borders between declensions, conjugations, and genders had become increasingly permeable in several morphophonological contexts (Sornicola, 2017: 85 ff.), without implying a change in meaning. Due to this inflectional flexibility, there is no reason to postulate new Early Medieval lemmas underlying the non-standard forms (Philippart de Foy, 2012).

Therefore, in *LLCT*, the new second-declension adjective *inanus* "void" (possibly reinforced by the second-declension *nanus* "dwarf", given that the form *nanis* is attested seven times in *LLCT1*) and the third-declension genitive/dative anthroponym *Ursoni* (genitive) "Ursus" with a Late Latin nasal declension are lemmatized under the corresponding standard lemmas: the third-declension *inanis* and the second-declension *Ursus*, respectively. This is done even though the ending *-us* is not etymologically derived from *-is* nor *-oni* from the standard genitive ending *-i*. A major subgroup of *LLCT* words with non-standard inflectional properties is formed by nouns which have undergone a gender change, such as *seculi* "centuries" in *super isti futuri seculi* "over the future centuries", where *seculi* with the masculine nominative plural ending *-i* is lemmatized under the standard Latin neuter *saeculum* (whose nominative plural is *saecula*) (Korkiakangas and Passarotti, 2011: 108). Likewise, *offertas* "offerings", seemingly a feminine accusative plural that had developed from the collective neuter plural in *-a*, *offerta*, is lemmatized under the Late Latin neuter singular lemma *offertum* (Adams, 2013: 431-432; Väänänen, 1981: 101-105).

An assignment of separate lemmas, such as *inanis* and *inanus*, to the above-mentioned standard and non-standard forms, respectively, would be a bad solution, not only because it ignores the historical development of Latin, but also because lemmas are by definition independent units of meaning, as stated in Section 1. In the cases above, inflectional change does not affect meaning. On the other hand, there are genuinely homonymous lexemes with different meanings that have to be lemmatized under separate lemmas

(Murphy, 2010: 84). An example of homonymous lemmas in English are (*to*) *lie* "(to) speak falsely" and (*to*) *lie* "(to) rest horizontally". They are sometimes registered under separate entries in English dictionaries, especially if they belong to different parts of speech, such as the above verbs and the noun *lie* "false statement".

In Latin, verb lemmas are rarely homonymous with lemmas of other parts of speech, contrary to English. Homonymous lemmas are potentially problematic for corpus linguistics, but in practice they are almost always disambiguated by their part of speech and syntactic properties. For example, the verb *intro* is inflected in person, tense, mood, and voice while the Late Latin preposition *intro* is indeclinable, and they have completely different distributions. An insuperable ambiguity only arises with lemmas such as *jus* "justice" vs *jus* "broth, juice", which are both nouns.

With *LLCT*, it is defined that homonymy arises when identical lemmas have different parts of speech or when they are etymologically of different origin. On the other hand, the evolutionary principle presented above entails that semantic differentiation does not give rise to new lemmas (Murphy, 2010: 87-90). For example, *band* "strip or loop of material" and *band* "musical group" in English would not be considered different lemmas in *LLCT* because they derive etymologically from the same origin. An opposite approach is seen, for example, in the *Longman Dictionary of the English Language* (Gay *et al*., 1984, eds.: 111), which gives the above nouns independent lemmas *[1]band* and *[3]band*.

Technically, the lemmatization of *LLCT1* follows the original *LDT* style in that homonymous lemmas are disambiguated by specifier numbers, for example *intro1* and *intro2*, with non-homonymous lemmas marked with *1* by default (e.g. *nomen1*). As was stated, the *LDT* lemmatization is based on the Perseus Dynamic Lexicon, which reproduces the entries of Lewis and Short (1879). Since Lewis and Short did not aim at keeping the lemmas at a minimum, the *LDT* style includes quite a number of cases with homonymous lemmas that could be subsumed under one lemma (e.g. *pecus* "cattle, beast" with three entries). Moreover, no clear distinction is made between past participles and homonymous nouns derived from them, such as *exitus* "gone out" and *exitus* "departure". Even *LLCT2* initially exploited specifier numbers, but in the current revised version of *LLCT2*, the numbers were removed and the ten remaining pairs of homonymous lemmas disambiguated by way of a specifier that usually indicates the part of speech, such as *latus^n*(*oun*) "side, flank" as opposed to *latus^a*(*djective*) "wide" and

intro^v(*erb*) as opposed to *intro^p*(*reposition*)[14]. This was done to respect the definition of a lemma as a semantically distinct unit, although in the case of *LLCT*, the use of specifiers is strictly speaking redundant, given that all the homonymous lemmas of *LLCT* can also be disambiguated by referring to the part-of-speech annotation layer. For the present, there are no genuinely homonymous lemmas in *LLCT*, such as the two nouns *jus*.

Having said all this, some borderline cases still remain in the lemmatization of *LLCT*. The word *locus* "place", originally a masculine, is very often used with the neuter endings *locum* and *loca*. The current version of *LLCT1* still lemmatizes forms with undeniable neuter endings under *locum1*, while the forms with endings that can be attributed to the masculine lemma go under *locus1*, contrary to the parsimony principle. In *LLCT2*, this incoherence has been corrected, and all forms are now lemmatized under *locus*. Likewise, in their current state, both *LLCT1* and *LLCT2* separate the lemmas *dominus* and *domnus*, although the latter clearly derives from the former. The lemma *dominus* "Lord" almost exclusively refers to God, while *domnus* "lord" is used as an appellation of human beings, e.g. *domnus Iacobus episcopus* "lord Jacobus, the bishop" (cf. Italian *don*). The treatment of *locus* has to be rectified in *LLCT1* and that of *domnus*/*dominus* both in *LLCT1* and *LLCT2* in pursuance of an anticipated general revision of *LLCT1*.

6. *Conclusion*

This paper has analysed the theoretical bases of the lemmatization of the Late Latin Charter Treebanks by discussing in detail the principles that were followed in their lemmatization: the evolutionary principle and the parsimony principle. In addition to the fact that no generally accepted guidelines for the lemmatization of Latin exist, the non-standard Early Medieval fea-

[14] The other homonymous lemmas marked with a specifier in *LLCT2* are *amicus^n*(*oun*) "friend" as opposed to *amicus^a*(*djective*) "friendly", not present in *LLCT*; *excepto^adv*(*erb*) "except" as opposed to *excepto^c*(*onjunction*) "except", *excepto^p*(*preposition*) "except (for)", and *excepto^v*(*erb*) "to exclude"; *finis^p*(*reposition*) "up to" as opposed to *finis^n*(*oun*) "end, region"; *intrinsecus^n*(*oun*) "indoor movables" as opposed to *intrinsecus^adv*(*erb*) "inwardly", not present in *LLCT*; *labor^n*(*oun*) "work" as opposed to *labor^v*(*erb*) "to glide", not present in *LLCT*; *papa^father* "father, pope" as opposed to *papa^pappa* "the word with which infants call for food" (Lewis and Short, 1879: *s.v. papa*), not present in *LLCT*; *partio^n*(*oun*) "part, portion" as opposed to *partio^v*(*erb*) "to share", not present in *LLCT* (*partio* may be a contamination of *portio* "portion" and *parte*(*m*) "part" or *partitio* "partition"); *super^p*(*reposition*) "over, above" as opposed to *super^adv*(*erb*) "over, above".

tures of charter Latin pose challenges to all levels of linguistic analysis, not least to lemmatization. Particularly, the highly frequent proper names with no canonized spelling in Latin are difficult to lemmatize consistently. Many of the most challenging names are of Germanic origin.

The central problem of the Latin of *LLCT* is how to use the analytical apparatus arising from Classical standard Latin to annotate forms and lemmatize words that do not exist in that standard. Because Early Medieval Latin never formed a written standard of its own, no description of its grammatical categories or its vocabulary is sufficiently solid to serve as the basis of morphological annotation or lemmatization, hence the adherence to the grammatical description of Classical standard Latin. In order to leap the gap between the attested non-standard forms and the existing standard, a principle called 'the evolutionary principle' was introduced. This principle reduces the linguistic variants provoked by language evolution to their standard Latin ancestors.

It is relatively easy to apply the evolutionary principle to Latin-based common names and other parts of speech which do have a standard Latin ancestor, while the lemmatization of forms that have no standard-Latin ancestor is more challenging. These latter are Late Latin neologisms or loans from other languages, mainly from Germanic ones, and they usually display a number of different spellings. The word's attestations in *LLCT* and in other sources, if available, are first carefully analysed and relevant lexicographical studies consulted. Subsequently, the (morpho)phonologically most plausible ancestor is either chosen between the attested forms or reconstructed on their basis.

Due to their special role in naming individuals, proper names tend to show more phonological erosion and less corrective normalization than other vocabulary and, therefore, their etymological origins become more readily blurred. This issue is pronounced in charters, where both anthroponyms and toponyms are frequent. Proper names with standard Latin ancestors are usually lemmatized with little uncertainty, while proper names with foreign, mainly Germanic, origin pose the biggest challenges to the use of the evolutionary principle: the Germanic names of *LLCT* almost never have obvious standard variants. The decision on the lemma is based on the frequency and the language-historical plausibility of the form. However, the *LLCT* lemma of a Germanic-based proper name is not a faithful reconstruction of the underlying Germanic word but rather an abstraction based on the attested Early Medieval Latin forms.

As charters are original documents and their Latin is highly irregular, the lemmatization, as well as the morphological and syntactic annotation, also have to take mistaken expressions into consideration. According to the evolutionary principle, functionally nonsensical semantic mistakes are not corrected in the lemmatization, just like functionally impossible mistaken morphology is annotated formally as it stands.

The other general principle applied to the lemmatization of *LLCT*, i.e. the parsimony principle, is introduced to avoid unnecessary proliferation of lemmas. The parsimony principle lumps under one lemma the forms that have the same meaning but have changed their inflectional properties. On the other hand, there are genuinely homonymous lexemes with different meanings that have to be lemmatized under distinct lemmas. Based on the evolutionary principle, identical lemmas are only considered homonymous in *LLCT* if they have different parts of speech and they are not of the same origin etymologically.

The scrupulous analysis of the above issues has shown that the lemmatization of *LLCT* is not as coherent as it should be. While the bulk of the lemmas of common nouns and other parts of speech can be trusted, the lemmatization of proper names would clearly benefit from a careful harmonization, hopefully realized in pursuance of a future revision of *LLCT1* and later a revision of *LLCT2*.

References

Adams, J.N. (2013), *Social Variation and the Latin Language*, Cambridge University Press, Cambridge.

Amelotti, M. and Costamagna, G. (1975), *Alle origini del notariato italiano*, Giuffrè, Milano.

Anderson, J. (2007), *The Grammar of Names*, Oxford University Press, Oxford.

Auernheimer, B. (2003), *Die Sprachplanung der karolingischen Bildungsreform im Spiegel von Heiligenviten*, K.G. Saur, München / Leipzig.

Bamman, D. and Crane, G. (2011), *The Ancient Greek and Latin Dependency Treebanks*, in Sporleder, C., van den Bosch, A. and Zervanou, K. (2011, eds.), *Language Technology for Cultural Heritage*, Springer, Berlin / Heidelberg, pp. 79-98.

Bamman, D., Passarotti, M., Crane, G. and Raynaud, S. (2007), *Guidelines for the Syntactic Annotation of Latin Treebanks*, v. 1.3. [available online at *nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf*].

Barsocchini, D. (1837), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 5, 2, Francesco Bertini, Lucca.

Barsocchini, D. (1841), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 5, 3, Francesco Bertini, Lucca.

Bartoli Langeli, A. (2006), *Notai: scrivere documenti nell'Italia medievale*, Viella, Roma.

Bertini, D. (1836), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 4, 2, Francesco Bertini, Lucca.

Brown, K. and Miller, J.E. (2013), *The Cambridge Dictionary of Linguistics*, Cambridge University Press, Cambridge / New York.

Cecchini, F.M., Korkiakangas, T. and Passarotti, M. (2020), *A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages*, in Calzolari, N., Béchet, F., Blache, Ph., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2020, eds.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, European Language Resources Association (ELRA), Paris, pp. 933-942.

ChLA^{1} = Chartae Latinae Antiquiores*: Facsimile-Edition of the Latin Charters Prior to the Ninth Century*, Bruckner, A., Marichal, R. *et al.* (1954-2001, eds.), Urs Graf Verlag, Olten / Dietikon / Zürich.

ChLA^{2} = Chartae Latinae Antiquiores*: Facsimile-Edition of the Latin Charters, 2nd Series: Ninth Century*, Cavallo, G., Nicolaj, G. *et al.* (1997-2019, eds.), Urs Graf Verlag, Dietikon / Zürich.

Costambeys, M. (2013), *The laity, the clergy, the scribes and their archives: The documentary record of eighth and ninth-century Italy*, in Brown, W., Costambeys, M., Innes, M. and Kosto, A. (2013, eds.), *Documentary Culture and the Laity in the Early Middle Ages*, Cambridge University Press, Cambridge, pp. 231-258.

Du Cange, Ch., Charpentier, D.P., Henschel, G.A.L. and Favre, L. (1883-1887, [1678^{1}]), *Glossarium mediae et infimae latinitatis* (édition augmentée), Niort, Paris.

Forcellini, E., Furlanetto, G. and De-Vit, V. (1858-1875), *Totius Latinitatis Lexicon*. Voll. 1-4, Typis Aldinianis, Pratii.

Francovich Onesti, N. (2000), *Vestigia longobarde in Italia (568-774): lessico e antroponimia*, Artemide edizioni, Roma.

Francovich Onesti, N. (2002), *The Lombard names of Early Medieval Tuscany*, in Boullón Agrelo, A.I. (2002, *ed.*), *Actas do XX Congreso Internacional de Ciencias Onomásticas*, Fundación Pedro Barrié de la Maza, A Coruña, pp. 1141-1164.

Francovich Onesti, N. (2010), *Indizi di sviluppi romanzi riflessi nelle voci germaniche e nei nomi propri*, in «Germanic Philology», 2, pp. 67-101.

Frank-Job, B. and Selig, M. (2016), *Early evidence and sources*, in Ledgeway, A. and Maiden, M. (2016, eds.), *The Oxford Guide to the Romance Languages*, Oxford University Press, Oxford, pp. 24-34.

Gaffiot, F. (1934), *Dictionnaire latin-français*, Hachette, Paris.

Gay, H., O'Kill, B., Seed, K. and Whitcut, J. (1984, eds.), *Longman Dictionary of the English Language*, Longman, London.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. and Bémová, A. (1999), *Annotations at Analytical Level: Instructions for annotators* [available online at *http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf*].

Korkiakangas, T. (2016a), *Morphosyntactic realignment and markedness change in Late Latin: Evidence from charter texts*, in «Pallas», 102, pp. 287-296.

Korkiakangas, T. (2016b), *Subject Case in the Latin of Tuscan Charters of the 8th and 9th Centuries*, Societas Scientiarum Fennica, Helsinki.

Korkiakangas, T. (2017), *Spelling variation in historical text corpora: The case of early medieval documentary Latin*, in «Digital Scholarship in the Humanities», 33, pp. 575-591.

Korkiakangas, T. (2018), *Spoken Latin behind written texts: Formulaicity and salience in medieval documentary texts*, in «Diachronica», 35, pp. 429-449.

Korkiakangas, T. (in press), *Late Latin Charter Treebank: Contents and annotation*, in «Corpora», 16, 2.

Korkiakangas, T. and Lassila, M. (2013), *Abbreviations, fragmentary words, formulaic language: Treebanking medieval charter material*, in Mambrini, F., Passarotti, M. and Sporleder, C. (2013, eds.), *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities*, Bulgarian Academy of Sciences, Sofia, pp. 61-72.

Korkiakangas, T. and Passarotti, M. (2011), *Challenges in annotating medieval Latin charters*, in «Journal of Language Technology and Computational Linguistics», 26, pp. 103-114.

Lewis, Ch.T. and Short, Ch. (1879), *A Latin Dictionary*, Clarendon Press, Oxford.

Longrée, D. and Poudat, C. (2010), *New ways of lemmatizing and tagging Classical and Post-Classical Latin: The* LATLEM *Project of the* LASLA, in Anreiter, P. and Kienpointner, M. (2010, eds.), *Proceedings of the 15th International Colloquium on Latin Linguistics*, Institut für Sprachwissenschaft der Universität Innsbruck, Innsbruck, pp. 683-694.

McGillivray, B. (2014), *Methods in Latin Computational Linguistics*, Brill, Leiden / Boston.

Murphy, M. (2010), *Meaning variation: polysemy, homonymy, and vagueness*, in Murphy, M. (2010, ed.), *Lexical Meaning*, Cambridge University Press, Cambridge, pp. 83-107.

Niermeyer, J.F., Van de Kieft, C. and Burgers, J.W.J. (2002, eds.), *Mediae Latinitatis Lexicon Minus* (revised edition), Brill, Leiden.

Philippart de Foy, C. (2012), *Lemmatiser un corpus de textes hagiographiques: enjeux et modalités pratiques*, in Biville, F., Lhommé, M.-K. and Vallat, D. (2012, *éds.*), *Latin vulgaire - Latin tardif IX. Actes du IX^e^ colloque international sur le latin vulgaire et tardif, Lyon, 2-6 septembre 2009*, Maison de l'Orient et de la Méditerranée 'Jean Pouilloux', Lyon, pp. 481-490.

Schiaparelli, L. (1929), *Codice diplomatico longobardo*. Vol. 1: *Fonti per la storia d'Italia 62*, Tipografia del Senato, Roma.

Schiaparelli, L. (1933a), *Codice diplomatico longobardo*. Vol. 2: *Fonti per la storia d'Italia 63*, Tipografia del Senato, Roma.

Schiaparelli, L. (1933b), *Note diplomatiche sulle carte longobarde*, in «Archivio storico italiano», 19, pp. 3-66.

Smith, J.Ch. (2011), *Change and continuity in form-function relationships*, in Maiden, M., Smith, J.Ch. and Ledgeway, A. (2011, eds.), *The Cambridge History of the Romance Languages*. Vol. 1: *Structures*, Cambridge University Press, Cambridge, pp. 268-317.

Sornicola, R. (2017), *La morfologia nominale: polimorfismo e polifunzionalismo nei sistemi di flessione*, in Sornicola, R., D'Argenio, E. and Greco, P. (2017, *a cura di*), *Sistemi, norme, scritture: la lingua delle più antiche carte cavensi*, Giannini, Napoli, pp. 85-134.

Väänänen, V. (1981, [1963[1]]), *Introduction au latin vulgaire*, Klincksieck, Paris.

Wright, R. (2000), *Latino e romanzo: Bonifazio e il Papa Gregorio II*, in Herman, J. and Marinetti, A. (2000, *a cura di*), *La preistoria dell'italiano: atti della Tavola Rotonda di Linguistica Storica (Università Ca' Foscari di Venezia, 11-13 giugno 1999)*, Niemeyer, Tübingen, pp. 219-229.

Treebanks

LLCT1 = https://zenodo.org/record/3633607#.XjU4lSNS9EY.

LLCT2 = https://zenodo.org/record/3633614#.XjU6zCN7lEY.

Timo Korkiakangas
Department of Languages
University of Helsinki
Unioninkatu 40
00014 Helsinki (Finland)
timo.korkiakangas@helsinki.fi